*Zootaxa* 3609 (6): 593–600
www.mapress.com/zootaxa/


**Article**

ISSN 1175-5326 (print edition)
**ZOOTAXA**
ISSN 1175-5334 (online edition)

http://dx.doi.org/10.11646/zootaxa.3609.6.5
http://zoobank.org/urn:lsid:zoobank.org:pub:E8AD9AD9-C6E4-4B73-BAA1-B25CE16F39E9

# Publishing large DNA sequence data in reduced spaces and lasting formats, in paper or PDF

ALEXANDRE PIRES AGUIAR
*Universidade Federal do Espírito Santo, Departamento de Ciências Biológicas, Av. Fernando Ferrari 514, Goiabeiras, Vitoria, ES, 29075–010, Brazil. E-mail: aguiar.2@osu.edu*

## Abstract

Scientific publications carry a practical moral duty: they must last. Along that line of thinking, some methods are proposed to allow economically and structurally viable publication of DNA sequence data of any size in printed matter and PDFs. The proposal is primarily aimed at contributing for preserving information for the future, while allowing authors to avoid information splitting and complement storage *ex situ*, that is, in server machines, outside the publication proper. The technique may also help to solve the impasse between the ICZN *Code* requirement that a new nomen be associated to diagnostic characters for the taxon *vs*. the phylogenetic definition of taxa, based on cladograms only: sequence data are characters, and can now be easily and comfortably included in taxonomic publications, with direct textual mention to their diagnostic sections. The compression level achieved allows the inclusion of all wanted DNA or RNA sequences in the same printed matter or PDF publications where the sequences are cited and discussed. Reduced font sizes, invisible fonts, and original 2D black & white and color barcodes are illustrated and briefly discussed. The level of data compression achieved can allow each full page of sequence data, or about 5000 characters, to be precisely coded into a color barcode as small as a square of 1.5 mm. A practical example is provided with *Taeniogonalos woodorum* Smith (Hymenoptera, Trigonalidae). Free software to generate publishable barcodes from txt or FASTA files is provided at www.systaxon.ufes.br/dna.

**Key words:** base pairs, COI, data compression, permanence, publication, RNA, taxonomy

## Introduction

The publication of DNA or RNA sequences is problematic in its essence, due to some practical reasons. First, such sequences are usually large, counting hundreds or thousands of base pairs for each specimen or taxon treated in a paper. If printed with regular font size, this will usually lead to many extra printed pages, which is both cumbersome and often prohibitively expensive for editors and authors. Second, such sequences are primarily meant to be read or processed by machines, not humans, so it makes little sense printing them as conventional text. In fact, several initiatives already try to solve the need to store sequences, the GenBank and the Barcoding of Life Database (BOLD) currently representing some of the most widely known.

Printed text, however, remains as the most stable form of publishing information aimed to last. It still represents the nearest we can get of providing an accessible and permanent format for information, with nearly two thousand years of proved efficiency. Such permanence derives both from the fact that physical copies are produced, and from the fact that, currently, thousands of identical copies are distributed and stored throughout the world. Server machines, on the other hand, are often few for each publication, and are clearly and immesurably more fragile or unstable than paper, both physically and logistically. It seems therefore obvious that online documents are still far from achieving a similar degree of permanence to that of paper publications.

The permanence of electronic *vs*. paper publishing is however a heated debate, and it is not the aim of this work to review such extensive discussion—while self-sufficient, all techniques proposed herein can also be advantageously used *in addition* to the currently available options for storing molecular data, with gains for all. The reader can however find more in depth discussions directly or indirectly related to the permanence of electronic publications and internet links in Dubois (2003, 2010, 2011), Dimitrova & Bugeja (2007), Carlos & Voisin (2009),

*Accepted by A. Dubois: 1 Nov. 2012; published: 4 Feb. 2013*

Löbl (2009), Welter-Schultes *et al*. (2009), Michel *et al*. (2010), and Anonymous [ICZN] (2012), among others. The aim of the present paper is more strictly technical, focusing to propose viable, space-efficient, long-lasting methods to include all wanted DNA or RNA sequences, no matter how long they are, in the very same printed matter or PDF publications where the sequences are cited and discussed.

## Proposed techniques

*Small font sizes*. Since the sequences *per se* do not need to be read as prose, the most obvious option to save page space is to use the simplest and smallest possible font for which physical printing remains clearly discernible. The question is, discernible to whom? For humans, this would generally mean a font type and size close to that illustrated in Fig. 1. For machines, font type and size are much less problematic, at least in theory. The key question seems therefore to be the following: which is the smallest possible *code* (alphabetic or otherwise), that can be printed on paper, that allows full and precise retrieval of the encoded information, for both humans and machines? Retrieving the same information from the PDF version of the publication would also be equally important, at least currently.

**FIGURE 1**. Text with 10 thousand base pair sequence (simulated), printed with a clean (sans-serif) and reasonably small font (Arial Narrow, font size = 3.5 pt), still readable with the unaided eye. As regular text in *Zootaxa*, this sequence would occupy over 2 pages.

If alphabetical characters are to be used, as in the IUPAC (Anonymous 1970) notation (also used in the FASTA format), the cleanest form that can be used is that of Fig. 2, with each font construed on a matrix of 4 × 5 pixels. Printing resolution achieved by most journals would already allow for great reduction in the corresponding printed text size, as demonstrated by Fig. 3. Whereas present day scanners and OCR software do work best with much larger fonts, retrieving the information from such texts—once accurately printed—would not pose a true challenge, neither now and much less in the future.

**FIGURE 2.** Simplest possible rendering of IUPAC notation as proportional fonts, at the pixel level. (ACGTU WSMKR YBDHV -).

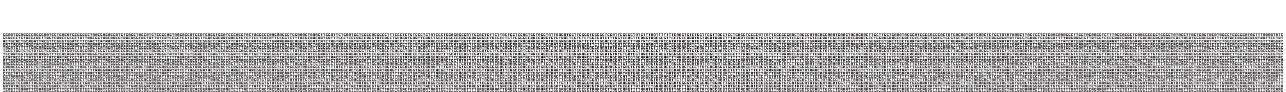

**FIGURE 3.** Ten thousand base pair sequence (same as in Fig. 1) printed as an image produced with the fonts of Fig. 2, at 300 dpi (printed matter resolution for *Zootaxa*).

*Monochrome barcode*. The next logical step is to devise even simpler, non-alphabetical characters, e.g., in the way of barcodes. The task is greatly simplified if only four characters, ACGT, need be coded. Uracil does not need

an extra coding scheme, sufficing to explain that coding was performed either for DNA or RNA. The simplest representation that can be achieved here is to use a couple of pixels in the states Black&White, so that A = W&W, C = W&B, G = B&W, and T = B&B. However, the white background could be confused as a series of W&W. To avoid this, the simplest measure is to state the total number of encoded bases (Fig. 4, top). A costly alternative is to mark the end of each code sequence with a black pixel, yielding therefore A = WWB, C =WBB, G = BWB, and T = BBB. A white background would immediately yield WWW, indicating the end of the coded sequence. The resulting barcode will be similar to that shown in Fig. 4 (bottom). Enlarging and interpreting both these kinds of black & white barcodes is a simple task, either via software or manually.

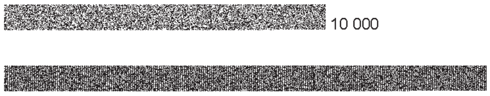


**FIGURE 4.** Ten thousand base pair sequence barcodes, based on two (top) or three (bottom) black & white pixels. Printed at 300 dpi.

*Color barcodes*. The double or triple pixels needed to represent each base can easily be reduced to a single pixel if colors are used. The closest technique currently in use is the "illustrative barcode", proposed by the BOLD initiative [http://www.barcodeoflife.org/]. It consists of vertical color-coded bars, separated from each other by white spaces, and representing the bases ACGT according to their electropherogram colors (Fig. 5).

**FIGURE 5.** BOLD's style "illustrative barcode", here representing the first 300 bp of Fig. 1.

This approach is however only partially useful in saving space, and may in fact take even more area than printing the raw sequence with small fonts. For example, by using the small font of Fig. 1, which is still readable with the naked eye, more than 1330 bp could be fitted in the same space here taken by Fig. 5, which codes instead for mere 300 bp—almost 4.5 times less. The illustrative barcode can however be reduced by half its length if the white spaces are removed. Smith *et al*. (2012), for example, published illustrative barcodes of similar length to that of Fig. 5, but representing 658 bp (see however item *Accuracy*, further below).

The compression rate can however be still greatly improved here if a single pixel is used instead of a bar, and if the smallest possible printing size on paper is taken into account. First, to maximize differences and accuracy of the printed colors, the color printing technique used by the target journal must be considered. Currently, there are two major techniques used worldwide: (1) color printing using CMYK pigments (Cyan/Magenta/Yellow/Black), which is the most common, and used also in most inkjet printers; and (2) color printing with RGB pigments (Red/Green/Blue), used by *Zootaxa* for example. The proportional combination of these three or four basic pigments on printed paper can emulate nearly all colors perceived by the human eye. Repeating precisely any particular color created on paper through mixed pigments, however, is obviously much more difficult than simply using the pure pigment. Thus, the CMYK pigments can be best used to directly represent the bases CATG, respectively – this particular order represents the closest approximation to electropherogram colors (C = blue, A = green, T = red, G = black). For presses using RGB pigments, the correspondence is precise: RGB+black = TACG. This is also the most distinct color set for the human eye (check for example Nathans 1999 and Rowe 2002).

Since a single pixel is now used for each base, the resulting 2D color barcode (Fig. 6A) has only 33.3% or 50.0 % of the size of its B&W equivalent, and is dramatically more compact than the illustrative barcode style. Where doubtful, expected nucleotides can be indicated by two preceeding and one subsequent white pixel (Fig. 6B), thus allowing the entire IUPAC notation to be coded with this method.


**FIGURE 6.** A. Ten thousand base pair sequence coded as a 4-color 2D barcode, printed at 300 dpi (printed matter resolution for *Zootaxa*). Color codes according to the RGB color space, as in electropherogram colors, as follows: A = green (RGB 0,255,0), C = blue (RGB 0,0,255), G = black (RGB 0,0,0) and T = red (RGB = 255,0,0). B. Diagram showing pixel-level encoding for GGAAT**R**GG**H**CCA**NN**. Note that R = either A/G, H = ACT (*i.e.*, not G), and N = any base (ACGT).

*Resolution of the Printed Barcode*. The smallest possible printed size (on paper) of the monochrome or color barcodes will also depend on the maximum printing resolution mechanically possible in the aimed journal. Color printing worldwide is often performed at 200–300 dpi (e.g., 300 dpi for *Zootaxa*), because this is already close or beyond the resolving power of the human eye. The size limit for the published barcodes on paper can thus be calculated simply by dividing the desired dimensions of the barcode, in number of pixels, by the printing resolution of the journal. For example, the barcode of Fig. 6 was generated as an image of 300 × 34 pixels. In the journal *Zootaxa* (max. resolution: 300 dpi), this is equivalent to a published image 300 pixels wide / 300 dots per inch = 1.0 inch, or 2.54 cm wide, and 34/300 = 0.1133 inches, or 2.9 mm high. This is the smallest possible printed size for that particular image where all individual pixels are still discernible on the paper pages of *Zootaxa*. If smaller than that, the printed dots (pixels) would overlap or mix, and the barcode would be ruined. Larger sizes are, conversely, increasingly more clear in representing each pixel independently.

All barcodes discussed above can be easily generated (see *Addendum*), and can be interpreted by equally simple software. Most importantly, however, they can all be quickly understood or cracked by any researcher in the future, independent of any machine or previous knowledge of how the barcode was generated. Note that a reading software to interpret the barcodes printed on paper is currently *not* needed, and would probably be of little use for the current generations of researchers. The discussed barcodes are aimed primarily as an economic, simple, machine-independent and long-lasting method for preserving sequence data information, all that within the same publication where they are discussed.

## Usage in PDFs

All options discussed above can also be fully reproduced in the corresponding PDFs, without any further difficulty. Specific properties of the PDFs, however, allow for other possibilities, and much greater compression, as discussed next.

*Wrapping*. For immediate retrieval of sequence data from PDFs, the smallest and simplest available font in the journal (e.g., Fig. 7) can be set to white, to make it invisible, and then wrapped behind the barcode image itself, or of course behind any other desired image, as done here in Fig. 8. Coloring the fonts white simply makes it unnecessary, in most cases, to format the text in a way that it can be fully covered by a given image.

**FIGURE 7.** Text with ten thousand base pair sequence (same as in Fig. 1), written with Arial Narrow, font size: 2 pt. (*Zootaxa* limit)

The hidden text still allows the user to select it, by selecting the image with the cursor, as would be done with regular text, and then copy and paste it in any text editor (Notepad is recommended, because it removes all formatting instantly, making text visible, and shown with regular font size), or even directly into DNA sequence formatting applications, such as http://informagen.com/Applets/Publish/, etc. If using a PDF, the reader can actually try it with Fig. 8 below. Text wrapping with images is performed by all major text or image editors, including free software.

*Full Color Barcodes*. The usage of full color 2D barcodes would promote the greatest possible compression of data, such as in Fig. 9. However, printing each color accurately *on paper*, and then reading them back with equal accuracy, are both likely to remain as intractable problems for a long time. Such coding system would also not be machine-independent, and definitely not easily broken without exact previous knowledge about the used standards on each case. Usage within PDFs is however fully possible: the trick is simply to associate unique sequences to unique color codes—note that the colors themselves are not important, because they cannot be accurately "viewed" or "read" by anything that is currently available. The true value of such a color barcode lays in its impressive capacity of storing information. The reduced size also allows it to be incorporated in regular pictures, as done here with Fig. 8 (upper margin of image).

**FIGURE 8.** *Taeniogonalos woodorum* Smith, lateral view, female (Hymenoptera, Trigonalidae). Image is wrapped above nucleotide sequences for 32 specimens (total of 20,617 bp), each preceeded by the respective sequence ID, as at boldsystems.org. Text in Arial Narrow 2 with line spacing 2 pt. (smallest possible setting in *Zootaxa*) and colored white. Albeit invisible, text can be selected, copied and pasted in any text editor (Notepad recommended). Color barcode at the upper margin of image encodes 68,225 nucleotides, corresponding to the sequences of all 121 specimens publicly availabe at *boldsystems* for this species (IUPAC notation, coded with RGBA for 13 bases per pixel; sequences separated by one white pixel; end of barcode marked by two white pixels). Image and data from Smith *et al*. (2012), used with permission.

**FIGURE 9.** Ten thousand base pairs, coded with CMYK for ACGT in sequences of 15 bases (Table 1). This is only a *fac simile*, because *Zootaxa* PDFs are encoded for RGB. Image not aimed for printed matter.

**TABLE 1.** Number of unique combinations (*UniqueComb.*) of coding letters for DNA for given sequence lengths (*Seq. length*), and the number of exceeding unique codes of RGB (*exc. RGB*) and CMYK or RGBA (*exc. RGBA*) in relation to the respective *UniqueComb*. Number *UniqueComb.* calculated for varied sequence lengths (6 to 16 bases long) considering also (1) only ACGT, (2) ACGTN, where N stands for any base, and (3) all possible coding situations, using the IUPAC (Anonymous 1970) notation, which includes 15 possible codes total (ACGT/WSMKRY/BDHV/N). Hyphen (-) used when number of DNA code combinations surpassed the number of possible color codes.

| Codes | Seq. length | Unique Comb. | exc. RGB | exc. RGBA |
|---|---|---|---|---|
| ACGT | 11 | 4,194,304 | 12,582,912 | 4,290,772,992 |
| ACGT | 12 | 16,777,216 | 0* | 4,278,190,080 |
| ACGT | 13 | 67,108,864 | - | 4,227,858,432 |
| ACGT | 14 | 268,435,456 | - | 4,026,531,840 |
| ACGT | 15 | 1,073,741,824 | - | 3,221,225,472 |
| ACGT | 16 | 4,294,967,296 | - | 0* |
| ACGTN | 10 | 9,765,625 | 7,011,591 | 4,285,201,671 |
| ACGTN | 11 | 48,828,125 | - | 4,246,139,171 |
| ACGTN | 12 | 244,140,625 | - | 4,050,826,671 |
| ACGTN | 13 | 1,220,703,125 | - | 3,074,264,171 |
| ACGTN | 14 | 6,103,515,625 | - | - |
| Full IUPAC set | 6 | 11,390,625 | 5,386,591 | 4,283,576,671 |
| Full IUPAC set | 7 | 170,859,375 | - | 4,124,107,921 |
| Full IUPAC set | 8 | 2,562,890,625 | - | 1,732,076,671 |
| Full IUPAC set | 9 | 38,443,359,375 | - | - |

(*) Same number as *UniqueComb.*; makes option unviable because codes for the white background, and for isolated ACGT at the end, are unavailable.

In computers, the component values for RGB, CMYK, etc, are often stored as integer numbers in the range 0 to 255, the range that a single 8-bit byte can offer. With the RGB color space, this means $256^3$ possible combinations, or 16,777,216 unique color codes. If the transparency component is added (RGBA), or if the CMYK color space is considered, the number jumps to $256^4$, or almost 4.3 billion unique codes. Table 1 summarizes some of the possibilities, indicating that each encoded pixel can store information equivalent to sequences of 6–15 bases. Converted to a PDF image, such as that of Fig. 9, this represents an impressive level of information compression

for scientific publications. For example, with CMYK encoding for sequences of 15 bases, the entire human genome, with about 3 billion bases, could be published in 33.7 PDF pages at *Zootaxa*. Using the regular font size for this journal, 600 thousand pages would be necessary.

The immediate advantage of using full color barcodes for storing information in PDFs—even considering that the printed version of such barcodes is useless—is that PDFs are server-independent, copies are easily multiplied, quickly becoming numerous, and end up stored in multiple places and devices. This is close to the advantages proportionated by printed matter, and clearly safer and more directly accessible than relying exclusively on server-based storage.

## Accuracy

For *any* of the approaches discussed above to work, information encoding and recovery must both be lossless, that is, 100% accurate. This requires, first of all, generating lossless image files, and then incorporating them into a PDF through a lossless storage procedure (see *Addendum*). Otherwise, the barcode will have only schematic or comparative value. In Smith *et al*. (2012), for example, the published barcodes can be quickly compared visually, but instead of only the four colors selected by the authors, the respective PDF file actually stored image files with 147,999 different CMYK colors and hues, making it rather challenging for any software to accurately interpret and recover all of the original sequences of bp. Data recovery in this case can still be achieved by visual inspection, but the process would be time-consuming, and prone to errors.

Note that, in order to recover all information from a barcode, the respective image must also be accurately *extracted* from the PDF, which is often not a simple "copy & paste" procedure (see *Addendum*). It is here important to appreciate that a PDF file usually stores an image as a separate object containing the raw binary data for the image. This is not an image in the sense of a JPG or a TIFF or a PNG file—it is the binary data for the pixels. As such, if the original information of each pixel is accurately saved into a PDF, then it will always be possible to retrieve the original information also accurately, independent of how accurately the image is *displayed* on the PDF or on a printed copy.

## Usage in taxonomy

The zoological *Code* (Anonymous 1999) requires that, in order to be available, a new taxon nomen (scientific name) be associated with statements of characters purported to differentiate the proposed taxon from closely related ones. Characters can be morphological, but also molecular, bioacoustic, cytogenetic, etc. A nucleic acid sequence can well be given as a diagnostic character that will make a new nomen available. This is in contrast with the so-called "phylogenetic definitions" of taxa, which are based only on the structure of a cladogram but do not mention characters (see also Dubois 2011: 44). A nomen published with only such a "definition" remains therefore nomenclaturally unavailable (Dubois 1999, 2006, 2007; Dubois *et al*. 2001; Ohler & Dubois 2012; Dubois & Raffaëlli 2012). Using sequences as diagnoses would be a good solution to this problem in many cases, but this is not a frequent practice precisely because printing complete sequences was a difficult issue until now. An immediate possibility would be to print the full sequence using one of the techniques proposed in this work, and then refer to the diagnostic parts of the sequence textually, thus attending the *Code* and making the new taxon nomen available.

## Remarks and acknowledgments

This paper was envisioned after a doubt about publishing full lists of sequences, launched by Rüdiger M. Schmelz (*Universidade da Coruña*, Spain) in a discussion list for *Zootaxa* subject editors, but all ideas proposed herein are, as far as the author is aware, fully original. The entire item "Usage in taxonomy" was kindly pointed out to the author by Alain Dubois; by including it in the paper, the author fully subscribes to it, and answers for its repercussion. Alain, in particular, and two anonymous reviewers, also contributed with other important suggestions and literature. David R. Smith (Smithsonian, USA) and co-authors kindly allowed the use of the image and

sequence data for Fig. 8. This research was supported by funding provided by *Fundação de Amparo à Pesquisa do Espírito Santo*/FAPES (Process 45.440.611/2009) and *Conselho Nacional de Pesquisas*/CNPq (Process 480507/2010-4), Brazil.

## References

Anonymous [IUPAC-IUB Commission on Biochemical Nomenclature] (1970). Abbreviations and symbols for nucleic acids, polynucleotides, and their constituents. *Biochemistry*, 9, 4022–4027. http://dx.doi.org/10.1021/bi00822a023

Anonymous [International Commission on Zoological Nomenclature] (1999) *International code of zoological nomenclature*. Fourth edition. London. International Trust for zoological Nomenclature, I–XXIX + 1–306.

Anonymous [International Commission on Zoological Nomenclature] (2012) Amendment of Articles 8, 9, 10, 21 and 78 of the International Code of Zoological Nomenclature to expand and refine methods of publication. *ZooKeys*, 219, 1–10. http://dx.doi.org/10.3897/zookeys.219.3944

Carlos, C.J. & Voisin, J.-F. (2009) A few remarks on the proposed amendment of the *International Code of Zoological Nomenclature* to expand and refine methods of publication. *Zootaxa*, 2198, 67–68.

Dimitrova, D.V. & Bugeja, M. (2007) The half-life of internet references cited in communication journals. *New Media & Society*, 9 (5), 811–826. [http://nms.sagepub.com/content/9/5/811.short; accessed 20 October 2012].

Dubois, A. (1999) Miscellanea nomenclatorica batrachologica, 19. Notes on the nomenclature of Ranidae and related groups. *Alytes*, 17 (1–2), 81–100.

Dubois, A. (2003) Editorial. Should internet sites be mentioned in the bibliographies of scientific publications? *Alytes*, 21 (1–2), 1–2.

Dubois, A. (2006) Naming taxa from cladograms: a cautionary tale. *Molecular Phylogenetics & Evolution*, 42, 317–330.

Dubois, A. (2007) Naming taxa from cladograms: some confusions, misleading statements, and necessary clarifications. *Cladistics*, 23, 390–402.

Dubois, A. (2010) Contributions to the discussion on electronic publication IV. Registration as a fourth floor of the nomenclatural process. *Bulletin of zoological Nomenclature*, 67 (1), 11–23.

Dubois, A. (2011) The *International Code of Zoological Nomenclature* must be drastically improved before it is too late. *Bionomina*, 2, 1–104.

Dubois, A., Ohler, A. & Biju, S. D. (2001) A new genus and species of Ranidae (Amphibia, Anura) from south-western India. *Alytes*, 19 (2–4), 53–79.

Dubois, A. & Raffaëlli, J. (2012) A new ergotaxonomy of the order Urodela Duméril, 1805 (Amphibia, Batrachia). *Alytes*, 28 (3–4), 77–161.

Löbl, I. (2009) Contributions to the discussion on electronic publication III. *Bulletin of zoological Nomenclature*, 66 (4), 307–308.

Michel, E., Nikolaeva, S., Dale-Skey, N. & Tracey, S. (ed.) (2010) Contributions to the discussion on electronic publication IV. *Bulletin of zoological Nomenclature*, 67 (1), 4–23.

Nathans, J. (1999) The evolution and physiology of human color vision: insights from molecular genetic studies of visual pigments. *Neuron*, 24, 299–312.

Ohler, A. & Dubois, A. (2012) Validation of two familial nomina nuda of Amphibia Anura. *Alytes*, 28 (3–4), 162–167.

Rowe, M.H. (2002) Trichromatic color vision in primates. *News in physiological Sciences*, 17 (3), 93–98.

Smith, D.R., Janzen, D.H., Hallwachs, W. & Smith, M.A. (2012) Hyperparasitoid wasps (Hymenoptera, Trigonalidae) reared from dry forest and rain forest caterpillars of Area de Conservación Guanacaste, Costa Rica. *Journal of Hymenoptera Research*, 29, 119–144. http://dx.doi.org/10.3897/JHR.29.3233

Welter-Schultes, F., Eikel, O., Feuerstein, V., Hörnschemeyer, T., Klug, R., Lutze, A., Tröster, G., Wieland, F., Willmann, R., Antezana Jerez, T., Baiocchi, D., Caldara, R., Núñez Cortés, C., Fenzan, W.J., Fery, H., Filmer, M., Gittenberger, E., Giusti, F., Horro González, J., Groh, K., Guerra, A., Hendrich, L., Jäch, M., Janssen, R., Jimenez Tenorio, M., Johanson, K.A., Kanase, Aruna A., Kenner, R.D., Koch, A., Lindner, N., Lorenz, F., Maehr, M.D., Manganelli, G., Martínez, S., Meregalli, M., Monteiro, A., Nielsen, S.N., de Oliveira, Á., Pearce, T.A., Pederzani, F., Petrov, P.N., Pola Perez, M., Poppe, G.T., Richling, I., Rolán, E., Sahlmann, B., Sama, G., Savage, J.M., Smetana, A., Stuardo, J., Sturm, C., Suárez Bustabad, M., Subai, P., Szekeres, M., Trigo, J.E., Tucker, J.K., van Vondel, B.J., Watts, C., Wiese, V. & von Wirth, V. (2009) Comment on the proposed amendment of articles of the *International Code of Zoological Nomenclature* to expand and refine methods of publication. *Bulletin of zoological Nomenclature*, 66 (3), 215–219.

**ADDENDUM**—Complementary details

*Pixels, dots, dpi, image files*. A pixel is an image unit in electronic devices. It is physically built with three tiny, independent light devices packed together – red, green, and blue lights, emulating the RGB system. The term dot is used for printed matter, because ink drops, unlike light devices, can be mixed on a same physical point: the dot. The smaller the dot a printer can render, or the smaller the size of the light devices, the more dots or pixels can be printed or displayed on a same area – both cases are measured as dpi units (= dots per inch, ≈ pixels per inch). Image *files*, however, store information for the pixels only, such as

"four red pixels" – these can then be printed or displayed under varied dpi, depending on the used rendering device and user choices. It makes no sense, therefore, to speak of dpi or image resolution for image files. The "dpi" information associated to some image files is a *separate* instruction added to the image file; it is interpreted by some devices (but not all), predefining how those pixels are to be displayed or printed.

*Software to generate the barcodes*. The Python(x,y) package, version 2.7.3.0, was used to write and run all codes. It is freely available at www.pythonxy.com. All original software used, with instructions, are freely available at www.systaxon.ufes.br/dna, mirrored at www.systaxon.com/dna. Both txt and FASTA formats are accepted as input files.

*RGB barcode for ACGT.* This is the quickest and cleanest choice for most demands. The Python code used to generate the respective barcode is provided below. To use it, install Python(x,y); copy and paste the code below on any text editor; save as txt; change the file extention from *.txt to *.py; create subdirectory C:\A\; place a txt file containing the desired DNA sequence in that subdirectory; click the py file to start; follow instructions on screen. The resulting image will be saved in C:\A\, as Plate.tif.

```
import Image

File = raw_input('File name: ') #txt files only; no paragraphs and no spaces allowed
Text = 'C:/A/%s.txt' % File #Reads in the txt file, which must be in subdirectory C:/A/
Data = open(Text).xreadlines()
ACGT = ''
for line in Data:
    for item in line:
        ACGT += item
ACGTnumber = len(ACGT)

A = Image.new('RGB', (1, 1), (0, 255, 0)) #Green
C = Image.new('RGB', (1, 1), (0, 0, 255)) #Blue
G = Image.new('RGB', (1, 1), (0, 0, 0))   #Black
T = Image.new('RGB', (1, 1), (255, 0, 0)) #Red

PlateWidth = input('Desired plate width (pixels): ')

Xlen = PlateWidth + 2
Ylen = (ACGTnumber / Xlen) + 2
Plate = Image.new('RGB', (Xlen, Ylen), (255, 255, 255))

coordX, coordY, step = 1, 1, 0
for base in ACGT:
    Plate.paste(eval(base), (coordX, coordY))
    coordX += 1
    step += 1
    if step == Xlen:
        coordX, step = 1, 0
        coordY += 1

Plate.save('C:/A/Plate.tif', dpi = (300, 300)) #Saves plate as 300 dpi TIFF file
```

*Lossless image compression and extraction in PDFs*. See for example http://www.binarynow.com/pdf-conversion/create-pdf-file-with-lossless-image-compression-for-high-resolution-printing/. For a couple of options with widely used software:

1. Adobe™Acrobat™ from MSWord: Save As → Adobe PDF → Options → check box "Create PDF/A-1a:2005 compliant file" → OK → Save

2. OpenOffice (text editor): File → Export as PDF... → in the "General" section check the box "PDF/A-1a"  Export

To extract an image from a PDF without information loss, the usual "copy & paste" is not recommended. This is because the displayed image is not necessarily an accurate interpretation of the original pixel data stored in the PDF. There are free software available for PDF image extraction (e.g., http://www.somepdf.com/downloads.html). Adobe™Acrobat™ will also accurately export PDF images.